Situation Awareness as a Predictor of Performance for En Route Air Traffic Controllers

Francis T. Durso, Carla A. Hackworth, Todd R. Truitt, Jerry Crutchfield, Danko Nikolic, and Carol A. Manning

Air traffic control instructors controlled simulated traffic while a variety of techniques for determining situation awareness (SA) were implemented. SA was assessed using a self-report measure (the Situation Awareness Rating Technique, or SART), a query method that removed information on the plan-view display (the Situation Awareness Global Assessment Technique, or SAGAT), a query technique that did not have a memory component (the Situation-Present Assessment Method, or SPAM), and the detection of errors integrated into the scenarios (implicit performance). We used these measures of SA together with a measure of workload, NASA Task Load Index (TLX), to predict two different performance measures: (1) an over-the-shoulder, subjective assessment by a subject matter expert (SME), and (2) a count of the number of control actions remaining to be performed at the end of the scenario. The SME evaluation was predicted by workload and the controller's appreciation of both the present and the future. The remaining-actions count (RAC) was predicted by the controller's appreciation of the future. In fact, an appreciation of the present led to poorer RAC scores: the better the participant was at answering questions about the present or the better he or she understood the present situation, the larger the number of actions that remained to be performed. The results have implications for the relationships among workload, SA, and performance, and suggest limitations on several of the measures currently pro-

Francis T. Durso, Carla A. Hackworth, Todd R. Truitt, Jerry Crutchfield, and Danko Nikolic are with the University of Oklahoma, Norman, OK. Carol A. Manning is with Civil Aeromedical Institute, Oklahoma City, OK.

Received April 1, 1997; accepted November 7, 1997.

Air Traffic Control Quarterly, Vol. 6(1) 1-20 (1998)
 CCC 1064-3818/95/030163-20

posed as SA techniques. The results confirm that future versus present is an important conceptual difference in ATC. More important, they suggest that a controller who remains overly focused on the present may do so at the expense of the future.

INTRODUCTION

The issue of situation awareness (SA) has presented quite a conundrum for applied investigators and basic researchers. Although SA has a number of theoretical definitions (e.g., Endsley, 1994; Fracker, 1988), most recognize SA as a cognitive construct distinct from workload (e.g., Endsley, 1993) but capable of impacting performance in a number of dynamic environments. For example, controlling air traffic is clearly a cognitive activity in a dynamic environment, and controllers recognize the value of maintaining good SA, or "the picture" as they call it.

If SA is neither performance nor workload, how can it be understood more precisely than "the picture" or more specifically than the cognitive component required to manage a changing environment? Intuitively, SA is the operator's understanding of the dynamic situation, including the current and likely future states. It includes knowing the situation in which one finds oneself when that situation has changed, what to do in the situation, what should follow from that situation, and how the situation relates to the operator's goals. An early, but specific, definition captures much of what is critical to SA: "the ability to envision the current and future disposition of both Red and Blue aircraft and surface threats" (Tolk and Keether, 1982, cited in Fracker, 1988:102). Endsley's (1988a:97) generalization, "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future," keeps the critical aspects of Tolk and Keether's definition while extending it beyond fighter aircraft. In both definitions, the distinction between the present and the future is highlighted.

SA is typically discussed as a characteristic of the operator in a particular environment. The environment is a dynamic one in which the operator has responsibilities or goals that impact the environment. It is this goal-directed aspect of SA that highlights the importance of future events. This focus on the future helps distinguish SA from other related cognitive constructs, such as understanding or perception. Although SA includes understanding and perception, it focuses on the future more than either of these other constructs. Durso et al. (1995a) found that comprehension of the current situation distinguished good chess players (master or intermediate) from bad players (novice), but could not distinguish among the good players. However, the ability to answer questions about the future of the game did distinguish master-level from intermediate-level players. Presumably, good players have a better understanding of the current state than do poor players, but expert players differ from intermediate players because of better representations of the future.

Our understanding of SA can advance without a commitment to any particular conceptual view of SA. In the social sciences especially, operational definitions of otherwise vaguely defined constructs have often been useful starting points from which consensus conceptual definitions have emerged. In fact, for SA, several researchers have advanced our understanding by defining it operationally. Specifically, researchers have used self-report, query methods, and implicit performance measures. One straightforward method, the Situation Awareness Rating Technique (SART) (Taylor, 1990) simply asks the operator for a judgment on a number of dimensions presumably related to SA. The Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988b) is an on-line query technique that taps an individual's recent memory of the situation. With SAGAT, information normally available to the operator is removed, and a question selected randomly from a battery of questions is presented to the operator. The more queries correctly answered, the better is the operator's SA.

In a related procedure, Durso et al. (1995a) asked participants to respond to SAGAT-like queries, but all information normally available to the participants remained in view. Instead of measuring percent correct, the Situation-Present Assessment Method (SPAM) uses response latency as the primary dependent variable. Although SPAM is procedurally similar to SAGAT, the two differ in interesting ways. In addition to not requiring a memory component, SPAM acknowledges that SA may sometimes involve simply knowing where in the environment to find a particular piece of information, rather than remembering what that piece of information is. For example, a controller need not store in memory the call sign of an aircraft, but good SA may require that he or she know where to find the call sign should communication with the aircraft be required. In fact, controllers are sometimes surprisingly poor at responding to SAGAT questions about information that would normally be visible to them (see Endsley and Rodgers, in this issue). Finally, some researchers (Sarter and Woods, 1991) have argued for a procedure that assesses implicit performance. With this procedure, an error is incorporated into an otherwise typical simulation, and the operator's SA is assessed by the speed and accuracy with which that error is detected and corrected.

In the current study, we attempted to determine which of these four SA procedures—SART, SAGAT, SPAM, and implicit performance—were able to predict the performance of en route air traffic controllers. For each regression model, we included a measure of workload to determine whether the measures of SA supplied anything beyond this venerable construct. If SA is a viable and measurable construct, individuals should vary in their levels of SA, and this variance should be useful in predicting performance. If SA differs from workload, it should have predictive value above and beyond that of workload.

METHOD

Site

This study was conducted at the Radar Training Facility at the Federal Aviation Administration's (FAA) Mike Monroney Aeronautical Center in Oklahoma City, Oklahoma. The facility is equipped with two radar training laboratories that allow for the simulation of en route traffic situations using the fictitious AERO Center airspace.

Participants

Twelve ATC instructors participated in the study. All participants were full performance level (FPL) controllers with an average of 18.8 years in ATC. The time as an FPL ranged from 4 to 29 years and averaged 11.6 years. The controllers had worked in the capacity of instructor for an average of 7.9 years (range .17 to 38 years[1]). Participants were familiar with the airspace, but unfamiliar with the scenarios employed.

Scenarios

All scenarios were developed in consultation with a subject matter expert (SME). A total of five 30 min scenarios were used. All scenarios contained a mix of general aviation, commercial, and military aircraft. Scenarios A and B were implicit performance scenarios (Sarter and Woods, 1991). Errors involving pilot readback, pilot nonconformance with ATC instructions, and data entry by the D-side controller were contrived by our SME and incorporated into these two scenarios. Confederates playing the roles of the pilots and other necessary personnel were supplied cue sheets indicating what errors to perform and when; errors were designed to occur at varied intervals in the two scenarios. The types of errors chosen were those most often implicated in actual operational errors (Durso et al., 1995b; Redding, 1992; Rodgers and Nye, 1993). Five errors were included in each scenario. However, one error from each scenario was not included for

[1] Because of the air traffic controllers' strike of 1981 and its consequences, a participant could have been an instructor for a period longer than he was technically an FPL.

scoring purposes because of difficulties in the timing of these errors and the accurate collection of data. Thus, implicit performance scores were based on four errors for each scenario. No error was scheduled to occur sooner than 2 min after the position relief briefing that began the scenario, or with less than 2 min remaining in the scenario, or within 1 min of another error.

Scenario A was designed for an individual performing both R- and D-side functions and contained a total of 21 aircraft: 7 arrivals, 7 departures, and 7 overflights. The four experimentally induced errors analyzed in Scenario A were pilot report of discrepant altitude, pilot readback error, nonconforming pilot, and failure to acknowledge instruction.

Scenario B was designed for an R- and D-side controller team. It contained a total of 29 aircraft: 7 arrivals, 10 departures, and 12 overflights. The four experimentally induced errors analyzed in Scenario B were D-side computer data entry error, pilot readback error, nonconforming pilot, and D-side controller prematurely suppressing the data block. The use of a confederate, D-side controller (SME) in Scenario B allowed the introduction of data handling errors. For example, the D-side controller entered and displayed a new but incorrect route (to Kansas City) on the radar screen.

Scenarios C, D, and E were designed for use during testing of the SART, SAGAT, and SPAM methodologies. These scenarios were to be controlled by an R-side controller only and were created to be approximately equal in complexity, as judged by our SME. Scenario C had 6 arrivals, 7 departures, and 7 overflights; Scenario D had 5 arrivals, 4 departures, and 11 overflights; and Scenario E had 1 arrival, 10 departures, and 15 overflights. No errors were built into these scenarios.

Performance Measures

SME Evaluations. The SME evaluated the controller's performance in Scenarios A, C, D, and E by observing his or her behavior; the SME's participation as D-side controller precluded the collection of SME evaluations during scenario B. The SME used the standard on-the-job training (OJT) evaluation form (FAA Form 3120-25). The SME indicated whether a set of specific behaviors was satisfactory, unsatisfactory, or in need of improvement, and also wrote comments about mistakes the controllers made during the scenarios.

Remaining Actions Count. Following each scenario, the SME determined the control actions that remained for each flight and completed a remaining-actions count (RAC) (Vortac et al., 1993). These actions reflect the behaviors necessary to move each flight successfully out of the controller's sector. Fewer remaining actions suggest more efficient control (e.g., Durso et al., 1995c; Vortac et al., 1993). That is, for any particular scenario, given the same starting configuration, a controller with fewer control actions remaining at the end of a specified time is viewed as having been more efficient in moving traffic.

Workload Measure

NASA Task Load Index (TLX). For the present experiment, we used a modified version of the NASA TLX form. NASA TLX (Hart and Staveland, 1988) is an instrument designed to assess several dimensions of workload, including mental demand, temporal demand, physical demand, effort, frustration, and performance. Participants were instructed to place an "x" on a line ranging from "low" to "high" on a scale from 0 to 96 mm, to reflect their perception of their workload during each of the scenarios.

Situation Awareness Measures

Query Techniques. With the assistance of the SME, Scenarios C, D, and E were examined, and six questions were designed to assess SA for each scenario. Three of the questions concerned the current situation (e.g., "Which has the lower altitude, TWA799 or AAL957?"), and three concerned a future situation (e.g., "Will DAL423 and FDX279 be traffic for each other, yes or no?"). Controllers were given a binary choice at the end of each question. The queries, appropriate presentation times, and viable foils were selected with the assistance of the SME. All questions were judged by the SME to address important information.

In most respects, our implementations of SPAM and SAGAT were similar. With both, one of the six questions was presented at the appropriate time, the controller answered the question, and the response was recorded. However, the two methods did differ in important respects.

With SPAM, the question was presented while all information normally available remained so. The SPAM question sequence began by activating the controller's landline. Participants were informed that all phone calls would come over a single landline, and further that some of the calls would come from "CAMI Center" which would query them about aspects of the situation. After the participant answered the landline, the experimenter read the question from a computer screen and initiated the timer. When the participant responded, the timer was stopped, and the experimenter recorded the response.

With SAGAT, a laptop computer was placed near the controller's work area on the side of the plan-view display (PVD) opposite the strip bay. When it was time for a question, the computer beeped, and the scenario was frozen. The participant immediately turned away from the PVD and toward the computer screen. The participant then read and answered the question by pressing the appropriate key, after which he or she returned to the primary task of controlling traffic.

Self-report Technique. The self-report technique used to assess SA was a version of SART. It included four scales: demand on attentional resources, supply of attentional resources, understanding, and SA. During the experiment, a tone was sounded, and the scenario was frozen. The controller turned from the screen and for each of the four scales placed an 'x' on a line that extended 0 to 51 mm. The scales were presented at the same points in the scenarios that would have queries in the SPAM and SGAT conditions.

Implicit Performance. In the individual version of the task (scenario A), the participants controlled traffic with the R-side and D-side positions combined. Participants were to control traffic as they would in the field while our SME observed to evaluate their performance. As noted, the SME measured controller performance using the OJT form. With the team version of implicit performance (scenario B), the participants were told that they would serve as R-sides as part of an ATC team. Our SME performed in the role of the D-side operator. Trained observers recorded reaction time in seconds from the occurrence of the error to the time the participant corrected the error. The observers listened to pilot-controller communications through headphones and recorded the reaction time via a laptop computer positioned behind the participants.

Design and Procedures

Participants controlled traffic across five air traffic scenarios. Thus, a within-subjects experimental design was used. All participants first completed an informed consent form and a biographical questionnaire. Prior to each scenario, participants were given appropriate instructions. Next, they were directed to their control positions and were provided with a position relief briefing from the SME. The briefing listed the equipment and operational conditions likely to be factors for the air traffic positions, and provided an overview of traffic patterns and VHF Omnidirectional Range (VOR) problems. The experiment was completed in two phases, with Scenarios A and B in the first half, and C, D, and E in the second. Following each scenario, participants completed the TLX workload measure.

Phase one comprised the two scenarios used as tests of implicit performance. These scenarios were used to assess the participants' ability to recognize and correct errors made by pilots and other members of the controller team in a timely manner. Scenario A was always the implicit performance—individual task; Scenario B was always the implicit performance—team task. The order of the two scenarios used for these tasks was counterbalanced. Following completion of the first phase, participants were interviewed about their experiences and opinions using a postexperimental questionnaire.

The second phase of the experiment involved the participants controlling traffic while completing various SA measurement instruments. Participants controlled traffic alone. The order of the three situation awareness methodologies—SAGAT, SPAM, and SART—was counterbalanced across the remaining three scenarios—C, D, and E. After each scenario, participants completed the modified TLX workload measure. Following completion of the second phase, participants were again interviewed about their experiences and opinions using a second post-experimental questionnaire.

RESULTS

In the following analyses it is important to keep in mind that the SART, SAGAT, and SPAM tasks were counterbalanced across three scenarios. Thus, differences among these measures cannot be attributed to inherent differences in the scenarios. However, the implicit performance tasks, by their nature, demanded that specific scenarios be designed for both the individual and team versions of the implicit performance task.

All multivariate analyses used the Wilk's Λ test statistic. All regressions used a stepwise procedure with an α of .15; all other analyses used an α of .05. Because of the relatively small number of participants (N = 12), shrinkage was addressed by reporting the adjusted R^2.

Performance Measures

Comparison of Scenarios. We began by comparing the five scenarios for each of the two performance measures—SME evaluations and RAC. SME evaluations using the FAA OJT form were tallied. A count of the number of less-than-satisfactory scores (i.e., "unsatisfactory" and "needs improvement") out of 27 categories was made.

~~Since the SME evaluations and the RAC are both performance~~ measures, they could have much in common. On the other hand, the two measures assess performance differently and may even focus on different aspects of the ATC task or on different components of SA. The SME evaluations are subjective, are performed by an individual skilled at the task, explicitly consider a myriad of task components, and are performed throughout the task (although the final check marks may occur at the end). The RAC index is objective, as argued earlier; considers task components only indirectly, and in fact may focus on different task components than the SME evaluations; and is distilled to the traffic situation at the end of the scenario.

A correlation of RAC and SME evaluations across the 12 participants was conducted separately for each of the four scenarios in which both measures were taken. The two performance measures were surprisingly unrelated. The correlations were −.05 (Scenario A), −.47 (Scenario C), +.14 (scenario D), and +.11 (Scenario E). These low or negative correlations suggest that the information captured by the RAC differs considerably from that reflected in the SME's evaluation. There are a number of reasons why these measures may differ, including the difference in subjectivity, the manner of data collection, and so on. However, as suggested in the later analyses, at least part of this difference is due to the fact that RAC is heavily dependent on the controller's appreciation of the future, whereas the SME evaluations depend on both present and future components.

Finally, we correlated SME evaluations from one scenario with those from another, and RACs from one scenario with those of another (see Table 1). The SME evaluation correlations tended to be quite high, with five of the six being significant. Part of the success here may lie in the fact that the SME is likely to impose additional consistency on the evaluations. The RAC intercorrelations were often more modest, with only four of ten showing any statistical significance. However, these correlations are also uniformly positive and sometimes quite substantial (e.g., $r = +.87$). Overall, Table 1 provides some evidence that individuals tended to maintain their relative standing in performance across the scenarios; that is, a good

Table 1. Intercorrelations Among the SME Ratings (Top) and RAC (Bottom) for the Five Scenarios

SME ratings could not be obtained in the Team (B) scenario.

Correlations	A	B	C	D	E
A					
SME		N/A	+.75*	+.70*	+.33
RAC		+.29	+.87*	+.19	+.53**
B					
SME			N/A	N/A	N/A
RAC			+.31	+.52**	+.62*
C					
SME				+.81*	+.63*
RAC				+.09	+.49
D					
SME					+.59*
RAC					+.15

*$p < .05$; ** $p < .10$.

controller in one scenario tended to be a good controller in the others. In general, this was true whether performance was measured by the SME or RAC, despite the differences between the two measures.

Workload Measures. TLX subscale scores were determined for each participant by measuring the distance from the low anchor to the participant's judgment point. With the exception of the performance subscale, the subscales of the TLX correlated highly and positively. Intercorrelations among mental demand, physical demand, temporal demand, and effort ranged from a low of +.87 to a high of +.95. Frustration correlated less well with these factors, but the correlations were still substantial, ranging from +.42 to +.68. Thus, a controller who viewed the task as mentally demanding also viewed it as physically demanding, temporally demanding, requiring a high level of effort, and relatively frustrating. Performance tended to correlate negatively with the other subscales, as would be expected. The high intercorrelations among the scales suggest that in subsequent analyses, such as the multiple regression analyses reported later, one subscale may enter the equation to the exclusion of its correlated neighbors, and which particular subscale it is may not matter. Overall, it appears that the TLX, at least as used here as a one-time, end-of-the-scenario measure, produces two important components—workload and subjective performance.

SA Measures

SART. SART scores were determined by measuring the distance (in millimeters) from the low anchor to the participant's judgment mark. The midpoint of each scale was 25.5, with a minimum of 0 and maximum of 51. The controllers indicated that they had an adequate supply of resources (M = 34), leading to good understanding (M = 44) and good SA (M = 45), for scenarios that they considered to be not very demanding (M = 20). The intercorrelations among the SART subscales were nonsignificant, with the exception that the SA subscale was positively and reliably correlated with understanding (r = +.88), suggesting that the controllers made little distinction between understanding and SA.

SPAM. Frequency and mean response latencies to future and present queries were computed, along with the mean time to answer the landline. Participants took almost 10 s to answer the landline and then took another 4 s to answer the query. As expected, the participants were quite accurate, especially if queried about the present situation. Response latencies were comparable for present and future queries. None of the SPAM intercorrelations reached conventional levels of significance.

SAGAT. Not surprisingly, percent correct scores for SAGAT were low as compared with SPAM. There was a moderate but nonsignificant correlation between percent correct for present and future queries. If one takes the perspective that future and present queries are merely two parts of an overall SAGAT score, then the +.35 correlation represents a rather poor split-half reliability. If, instead, one takes the perspective that future and present queries capture two important but orthogonal dimensions of SA, the correlation provides mild support for this thesis.

Implicit Performance. Number of errors detected and latency in making a detection were recorded. If an error was never detected, it contributed no datum to the latency analyses. Subjects noticed as many errors when assisted by a D-side controller (M = 54 percent, range = 25–100 percent) as when controlling traffic alone (M = 50 percent, range = 25–75 percent). In addition, controllers who did relatively well under the single-staffing condition did not necessarily do well under the team-staffing condition, as indicated by the small, nonsignificant correlation (r = −.18) between the number of errors identified in the two conditions.

Prediction of Remaining Actions and Expert Evaluations

These sets of analyses explored the ability of each of the SA procedures to predict the performance measures corresponding to that scenario. For example, we attempted to use SART and TLX taken during the SART scenario to predict the RAC and the SME evaluations for the SART scenario.

Given that the RAC and SME evaluations were surprisingly unrelated, it is not at all obvious how models developed for predicting SME evaluations should compare with models developed for predicting RACs.

The following analyses reveal which aspects of the SA measures contribute to predicting performance beyond any contribution by workload. The SA contributions reflect possible differences in both subjects and scenarios. Interpretations of the regressions should not assume that the predictive value of an SA measure is due solely to, for example, differences in the controller's SA abilities. Significant SA predictors are able to detect differences in individuals, scenarios, or both.

Prediction of SME Evaluations. The regression analyses for the SME evaluations appear in Table 2. SART was successful in predicting SME evaluations. The SART supply subscale combined with the TLX Mental demand subscale ($p < .06$) to account for 35 percent of the variance in SME evaluations. Low perceived supply and high

Table 2. Regression Summaries Predicting SME Evaluations

	Workload		SA		
	β weight	Variable	β weight	Variable	Adjusted R^2
Implicit Performance (Individual)		Temporal demand		Errors detected	.69
SART	.0099**	Mental demand	−.0270*	Supply	.35
SAGAT			−.0102**	Future queries	14
SPAM	−.0137**	Mental demand	−.0358*	Present queries	.53

*$p < .05$; **$p < .15$

perceived mental demand led to a poorer evaluation (cf. Selcon et al., 1991).

SAGAT had limited success at predicting SME evaluations. The more queries about the future a controller answered correctly, the better was his or her SME evaluation ($p < .13$), accounting for 14 percent of the variance.

SPAM was successful in predicting SME evaluations. A model including the number of present questions answered correctly and the TLX mental demand subscale ($p < .02$) predicted 53 percent of the variance in the SME evaluations. As with SAGAT, the more questions answered, the better evaluated was overall performance. However, in the SPAM analysis the critical questions were present-oriented. Finally, in contrast with other appearances of mental demand (e.g., SART analysis), here low mental demand implied more negative comments by the SME. Because low mental demand may suggest good performance or bad, this subjective workload component appears to be an unreliable predictor of SME evaluations.

Finally, implicit performance was able to predict SME evaluations. In this case, temporal demand from the TLX and the number of errors detected predicted 69 percent of the variance ($p < .003$). Greater perceived temporal demand led to poorer performance evaluations, and the fewer errors detected, the poorer were the performance evaluations.

Overall, SME evaluations were predictable by a combination of workload and SA measures. Having a high supply of resources, answering both future and present questions correctly, and detecting errors incorporated into the scenarios led to better SME evaluations.

Prediction of Remaining Action Counts

Regression analyses for RAC are summarized in Table 3. Of the SART subscales, demand and understanding combined to predict

Table 3. Regression Summaries Predicting RAC Evaluations

	Workload		SA		
	β weight	Variable	β weight	Variable	Adjusted R^2
Implicit Performance (Individual)		Performance			.29
Implicit Performance (Team)	No	variable	entered	the model	N/A
SART			.4585*	Demand	.27
			.7026**	Understanding	
SAGAT	−.1733*	Effort	−.1513*	Future queries	.74
			.2365*	Present queries	
SPAM			.2917**	Reaction time (future)	13

*p < .05; **p < .15

RAC ($p < .11$) and accounted for 27 percent of the variance in RACs. The demand factor is easily interpreted: the greater the overall demand perceived by the participant, the more control actions remained to be performed. However, the understanding factor is not easily interpreted, because the model indicates that understanding and RAC are positively correlated. In other words, the more understanding professed by the controller, the more actions remained to be performed at the end of the scenario. One obvious explanation is that these controllers were not very good at reflecting on their understanding, and thus subjective measures of SA may be inappropriate in the ATC environment. On the other hand, some of the other measures also raised similar concerns, and so we will return to an alternative interpretation of the understanding effect after considering the other analyses.

SAGAT future and present queries combined with the TLX effort subscale ($p < .004$) to account for an impressive 74 percent of the variance in RACs. Again, part of the model is easily interpreted: the fewer future questions answered correctly, the more actions remained to be performed. Also, the less perceived effort required, the better the participant performed. However, the better the participant was at answering questions about the present situation, the more actions remained to be performed at the end of the scenario. Because the raw correlations between the SAGAT factors and RAC were of opposite signs (ruling out a suppresser effect), we explored this result further by classifying participants as poor (0 or 1 correct) or good (2 or 3 correct) on the two types of questions–present and future. This classification yielded participants who did well on both (good SA), poorly on both (poor SA), well on future but not present (future-focused style), and well on present but not future (present-focused style). The present-focused (N = 4) controllers had the poorest performance, with 24 remaining actions; the future-focused (N = 2) controllers had the best RAC performance, with only 8 remaining actions.

This aspect of the SAGAT results is reminiscent of the SART understanding results and may suggest that the more one focuses on or understands the present situation, the poorer one will score on a measure of efficiency such as RAC. Assuming that "understanding" in SART is interpreted to mean understanding the present, a similar explanation can be applied to that analysis.

SAGAT's success at predicting RAC is substantial, but it indicates that some queries may be positively related to variables of interest, and others may be negatively correlated. For example, imagine a battery of SAGAT queries that focused on the present; we might find that individuals who did poorly on SAGAT actually performed better on the task or actually had better SA of impending events. Thus, the current data suggest that query techniques can be improved if greater control is exercised over the types of questions asked. The current study and previous work (Durso et al., 1995a) suggests that future versus present is an important difference.

SPAM generated a one-factor model predicting RAC. The time required to answer a query about the future ($p < .14$) accounted for 12 percent of the variance in RACs. Consistent with SAGAT, SPAM indicated that the slower the participants were to respond to questions about the future, the more actions remained to be performed at the end of the scenario.

Finally, for implicit performance, TLX performance ($p < .05$) entered the model, accounting for a respectable 30 percent of the variance, but no implicit performance measure contributed beyond the controller's perceived level of performance.

Overall, the analyses of RAC scores suggest that the control actions remaining at the end of the scenario were strongly dependent on the controller's ability or tendency to consider the immediate future during that scenario. This inference is indicated by the results from SAGAT and SPAM, both of which suggest that controllers who answer more future queries (SAGAT) or answer them more quickly (SPAM) will have fewer remaining actions at the end of that scenario. These analyses also suggest that controllers who focus instead on the present situation will perform poorly on the RAC measure. Controllers who could answer more present queries (SAGAT) or who understood the situation (SART) actually had more remaining actions at the end of the scenario. Presumably, controllers interpret the SART understanding of the situation to mean understanding of the present situation.

Prediction of Implicit Performance. The design of the current study allowed us to conduct an additional analysis—predicting implicit performance from the other SA measures. If SA is a unitary construct, a good measure of SA should capture the ability of participants to detect errors. We chose to predict implicit performance from the other SA measures for a number of reasons. Most views of SA would acknowledge that the ability to detect errors is a characteristic of good SA. However, the pragmatic aspects of using implicit performance require painstaking design of simulations, usually in consultation with a SME, and the amount of data collected is often small, making it difficult to reach conclusions backed by any statistical power. If a simpler method of assessing SA could be developed (e.g., SART, SAGAT, SPAM), it would have a great deal of practical value. Thus, a secondary purpose of this analysis was to determine whether a simple procedure could be developed within the ATC environment that could substitute for implicit performance measures.

Separate regressions were attempted for the individual and team implicit performance tasks. We expected that predictability across scenarios, as is the case here, would be lower than predictability within scenarios, as was the case in the performance analyses. Nevertheless, the results were disappointing. None of the SA measures were able to predict the number of errors correctly detected in the individual case. For the team case, SART failed to produce a model capable of predicting error detection. One encouraging finding came from SAGAT, which was able to predict 20 percent of the variance in error detection for the team situation. The only factor in the regression was the number of present questions answered correctly. The controllers who were especially adroit at answering present questions in one scenario tended to be those who were best at detecting the errors incorporated into a different scenario ($p < .08$). SPAM produced a model in which the time to answer the landline accounted for 33 percent of the variance. The longer the controllers took to answer the landline in the SPAM condition, the more errors they had detected in the earlier scenario ($p < .03$). If being present-oriented is predictive of errors, as SAGAT suggests, the longer landline times could be taken as an indication that present-oriented controllers are more reluctant to divert their attention in order to answer the landline.

DISCUSSION

The results presented above indicate that SA measures are able to predict performance beyond the predictability provided by workload. Both SME and RAC measures of performance were predictable from

SA measures. All SA measures were of some value in predicting the SME evaluations. Both an appreciation of the present and an appreciation of the future were useful predictors of SME evaluations. Only SAGAT and SPAM, the two query methods, had any predictive value for RACs. Implicit performance supplied nothing beyond that provided by perceived workload, and SART predictions were the opposite of what one would expect.

Why did SART and implicit performance measures have difficulty predicting RAC? One possibility is that both of these SA measures focus primarily on the current situation, ignoring the future component of SA, a component which is apparently critical to the RAC measures. The participants found the SART SA question to be virtually indistinguishable from the understanding question. In turn, the understanding question seems to have been interpreted as understanding of the current situation. Controllers who professed a greater understanding of a particular situation did poorly on the future-oriented RAC measure. Similarly, implicit performance may lack a future component. Several facts point in this direction. First, implicit performance was unable to predict RAC, a performance measure that was predictable by future-oriented but not present-oriented SA measures. Second, prediction of implicit performance depended on present-oriented factors, such as the present questions from SAGAT. Thus, error detection may depend primarily on the present component of SA. Although an error may have consequences for the future, in some sense it is available for detection in the present. It is an interesting methodological question whether errors can be constructed that emphasize the future component of SA, or whether all errors, regardless of their future impact, are detected with equal ease "in the present." According to the current study, however, implicit performance seems present-oriented, RAC future-oriented, and SME evaluations a little of both.

Perhaps the most interesting finding is that an appreciation of the present had effects opposite to those of an appreciation of the future, suggesting that controllers may attend to the present at the cost of the future. Performance, as assessed by RAC measures, was not merely unaffected by the present—it was actually poorer when an appreciation of the present was higher. Greater understanding (SART) and correct responses about the present (SAGAT) both appeared to hurt RAC performance. Recognizing that an appreciation of the present and future can have opposite effects on performance complicates all of the SA measures. For example, in the typical procedure of sampling randomly from a pool of questions, one must take into consideration the fact that a sample of questions dealing solely with the present situation can lead to a different evaluation of a system or individual operator than would a sample of questions dealing solely with the future situation. It is not merely that future and present questions capture different components of SA, but that they may be, at least for ATC, conflicting activities. A controller who focuses attention on the present during a particular scenario, and thus answers many such questions correctly, may well prove to be less efficient than a controller who answers fewer such questions correctly.

The current study was successful in pointing out the value of an appreciation of the future. It also supplied evidence that comprehension of the current situation and projection into the future are distinguishable and important components in the SA of air traffic controllers. The present and future may, however, lead to opposite effects on performance.

ACKNOWLEDGMENTS

This research was supported by contract DTFA-02-93-D-93088 from the FAA to Francis T. Durso. Requests for reprints should be sent to Frank Durso, Department of Psychology, University of Oklahoma, Norman, OK 73019, or e-mail: fdurso@ou.edu.

Thanks to Scott Gronlund, Mark Rodgers, and Dave Schroeder for comments on an earlier version of this work. We would like to thank the following personnel for their assistance in the completion of this project. Our subject matter expert (SME), Henry Mogilka, provided invaluable support. Through his efforts all scenarios were created and modified as needed with the help of Wayne Guthrie and Jim Ebeling. Henry scheduled instructors to participate in the study and assisted with the scheduling of the ghost pilots. Finally, he was very helpful with every request and evidenced consistent support. A special thanks to Rick Larson for authorizing flexibility in Henry's schedule so that he could meet our demands.

In addition, we are grateful to Dick Pollock for the use of the RTF facilities. We would like to recognize Betty (Zeke) Holmes for her efforts in arranging the necessary ghost pilots. The pilots who helped with this project were Joe Allen, Mike Evans, Russell Glazner, Scott Hughes, Bob Hutchinson, Connie Leiman, David Mitchell, and Sue Ruby.

Finally, we are very thankful to those individuals who volunteered to participate: Mark McKinney, Faith Arnell, Bob Bescancenecy, Skip Foster, Frank Wrisinger, A. J. Rotter, Ken Shaver, Bob Garcia, Mark Anderson, Carol Might, Mark Stemple, Paul DeBenedittis, and Peri Bennet.

ACRONYMS

D-Side	data-side controller
FAA	Federal Aviation Administration

FPL	full performance level
NASA	National Aeronautics and Space Administration
OJT	on-the-job training
PVD	plan-view display
RAC	remaining action count
R-Side	radar-side controller
SA	situation awareness
SAGAT	Situation Awareness Global Assessment Technique
SART	Situation Awareness Rating Technique
SPAM	Situation Present Assessment Method
SME	subject matter expert
TLX	Task Load Index
VOR	VHF Omnidirectional Range

REFERENCES

Durso, F. T., Truitt, T. R., Hackworth, C. A., Crutchfield, J. M., Nikolic, D., Moertl, P. M., Ohrt, D., and, Manning, C. A. (1995a). "Expertise and Chess: Comparing Situation Awareness Methodologies," in D. Garland and M. R. Endsley (eds.), *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*, pp. 295–303.

Durso, F. T., Truitt, T. R., Hackworth, C. A., Crutchfield, J. M., Ohrt, D. D., Hamic, J. M., and Manning, C. A. (1995b). "Factors Characterizing En Route Operational Errors: Do They Tell Us Anything About Situation Awareness?" in D. Garland and M. R. Endsley (eds.), *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*, pp. 189–195.

Durso, F. T., Truitt, T. R., Hackworth, C. A., Albright, C. A., Bleckley, M. K., and Manning, C. A. (1995c), *Reduced flight progress strips in en route ATC mixed environments*. Technical report, Cognitive Processes Laboratory, University of Oklahoma. FAA Technical report pending government review.

Endsley, M. R. (1988a), "Design and Evaluation for Situation Awareness Enhancement," in *Proceedings of the Human Factors Society 32nd Annual Meeting*, 1, Santa Monica, CA: Human Factors Society, pp. 97–101.

Endsley, M. (1988b), Situation Awareness Global Assessment Technique (SAGAT), in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference-NAECON*, 3, New York: Institute of Electrical and Electronics Engineers, pp. 789–95.

Endsley, M. R. (1993), "Situation Awareness and Workload: Flip Sides of the Same Coin," in R. S. Jensen and D. Neumeister (eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology*, Columbus, OH: Department of Aviation, The Ohio State University, pp. 906–11.

Endsley, M. R. (1994), "Situation Awareness in Dynamic Human Decision Making: ~~Theory," in R. D. Gilson, D. J. Garland, and J. M. Koonce (eds.), *Situational*~~ *Awareness in Complex Systems*, Daytona Beach, FL: Embry-Riddle Aeronautical University Press, pp. 27–58.

Fracker, M. L. (1988), "A Theory of Situation Assessment: Implications for Measuring Situation Awareness," in *Proceedings of the Human Factors Society 32nd Annual Meeting*, Santa Monica, CA: Human Factors Society, pp. 102-6.

Hart, S. G. and Staveland, L. E. (1988), "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in P. A. Hancock and N. Meshkati (eds.), *Human Mental Workload*, Amsterdam: North-Holland, pp. 139–83.

Redding, R. E. (1992). Analysis of Operational Errors and Workload in Air Traffic Control, in *Proceedings of the Human Factors Society 36th Annual Meeting*, Santa Monica, CA: Human Factors Society, pp. 1321–25.

Rodgers, M. D. and Nye, L. G. (1993), "Factors Associated with the Severity of Operational Errors at Air Route Traffic Control Centers," in M. D. Rodgers (ed.), *An Examination of the Operational Error Database for Air Route Traffic Control Centers*, DOT/FAA/AM-93/22, Washington, D. C.: Office of Aviation Medicine, pp. 11–25.

Sarter, N. B. and Woods, D. D. (1991), "Situation Awareness: A Critical but Ill-Defined Phenomenon," *The International Journal of Aviation Psychology*, 1, 45–57.

Selcon, S. J., Taylor, R. M., and Koritas, E. (1991), "Workload or Situational Awareness? TLX vs SART for Aerospace Systems Design Evaluation," *Proceedings of the Human Factors Society*, 35, 62–66.

Taylor, R. M. (1990), "Situation Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design," in AGARD-CP-478, *Situation Awareness in Aerospace Operations*, Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development, pp. 3–1 to 3–17.

Vortac, O. U., Edwards, M. B., Fuller, D. K., and Manning. C. A. (1993), "Automation and Cognition in Air Traffic Control: An Empirical Investigation," *Applied Cognitive Psychology*, 7. 631–51.

BIOGRAPHIES

Frank Durso is Professor of Psychology and Director of the Human-Technology Interaction Center at the University of Oklahoma. He is senior editor of the Handbook of Applied Cognition, to be published by Wiley. The focus of Dr. Durso's aviation-related interests has been on the cognitive processes of en route air traffic controllers.

Carla A. Hackworth, M.A., is a graduate student in social psychology at the University of Oklahoma. Her interests within aviation include expertise in air traffic control. Currently, she is working in the area of social intelligence with an interest in academic and industrial applications.

Todd R. Truitt, M.S., is a graduate student in cognitive psychology at the University of Oklahoma. His research interests include aviation, particularly air traffic control, expertise, and the relationship between interest and knowledge. Todd has served as an Engineering Research Psychologist Student Intern at the FAA William J. Hughes Technical Center, and he is also a private pilot.

Jerry M. Crutchfield is a third-year graduate student in psychology at the University of Oklahoma. His research interests are situation awareness and the relationship among cognitive interest, prior knowledge, and recall.

~~**Danko Nikolic** is a graduate student from Croatia. He received an M.S. degree~~ from the University of Oklahoma and is currently working on his Ph.D. His research interest is in mathematical modeling of various aspects of human cognition.

Carol Manning, Ph.D., is an Engineering Research Psychologist in the Human Factors Research Laboratory at the FAA's Civil Aeromedical Institute (CAMI), located in Oklahoma City, OK. She has been with CAMI since 1983, and is involved in the development of objective measures of task load and performance for the ATC specialist, using available National Airspace System data. Before joining the FAA, she worked for a year at the American Institutes for Research. She has a Ph.D. in Experimental Psychology from the University of Oklahoma (awarded in 1982), with an emphasis in decision theory.